Can multisensory training aid visual learning? A computational investigation

Robert A. Jacobs
Department of Brain and Cognitive Sciences, University of Rochester, Rochester, NY, USA

Chenliang Xu
Department of Computer Science, University of Rochester, Rochester, NY, USA

Although real-world environments are often multisensory, visual scientists typically study visual learning in unisensory environments containing visual signals only. Here, we use deep or artificial neural networks to address the question, Can multisensory training aid visual learning? We examine a network's internal representations of objects based on visual signals in two conditions: (a) when the network is initially trained with both visual and haptic signals, and (b) when it is initially trained with visual signals only. Our results demonstrate that a network trained in a visual-haptic environment (in which visual, but not haptic, signals are orientation-dependent) tends to learn visual representations containing useful abstractions, such as the categorical structure of objects, and also learns representations that are less sensitive to imaging parameters, such as viewpoint or orientation, that are irrelevant for object recognition or classification tasks. We conclude that researchers studying perceptual learning in vision-only contexts may be overestimating the difficulties associated with important perceptual learning problems. Although multisensory perception has its own challenges, perceptual learning can become easier when it is considered in a multisensory setting.

Introduction

Real-world environments are multisensory—they provide observers with visual, auditory, haptic (active touch), olfactory, and other signals conveying information about their underlying structures. Despite this, visual scientists typically study visual learning in environments that do not include signals from other modalities. Although this approach may simplify the study of some aspects of visual learning, it ignores other aspects that are likely to be essential to a comprehensive understanding. Indeed, more than 300 years ago, Bishop George Berkeley (1709/1910) speculated that infants acquire aspects of visual perception by correlating visual sensations with sensations arising from motor movements. A famous quote from Berkeley's book is "touch educates vision." More recently, Piaget (1952) used similar ideas to explain how children learn to interpret and attach meaning to retinal images based on their motor interactions with physical objects.

In this paper, we use a neural network known as a β variational autoencoder (β-VAE) to address the question, Can multisensory training aid visual learning? An advantage of β-VAEs is that it is easy to adjust their information capacities, meaning that the internal representations they acquire can be weakly, moderately, or strongly constrained. We use β-VAEs to study the internal representations that a learning system possesses when it is exposed to visual signals regarding the shapes of objects. These representations are studied in two conditions: (a) when the network is initially trained with both visual and haptic signals, and (b) when it is initially trained with visual signals only.

We find that visual-haptic training can lead to more abstract object representations that include, for example, information regarding the categorical structure of objects. Visual-haptic training can also lead to representations that are more viewpoint insensitive. We conclude that researchers studying perceptual learning in vision-only contexts may be overestimating the difficulties associated with important perceptual learning problems. Although multisensory perception has its own challenges, perceptual learning can become easier when it is considered in a multisensory setting (Shams & Seitz, 2008).

Background

It is only recently that vision scientists have begun to experimentally evaluate the idea that "touch educates

Citation: Jacobs, R. A., & Xu, C. (2019). Can multisensory training aid visual learning? A computational investigation. *Journal of Vision*, *19*(11):1, 1–12, https://doi.org/10.1167/19.11.1.

https://doi.org/10.1167/19.11.1 Received December 13, 2018; published September 3, 2019 ISSN 1534-7362

vision." Held et al. (2011) tested treatable, congenitally blind individuals. After sight restoration, it was found that these individuals were unable to visually match an object to a haptically sensed sample, though this ability developed within days suggesting that visual-haptic matching requires visual-haptic experience. Other researchers have studied the affects of visual-haptic experience on visual perception when vision and haptics are discrepant or when visual information is highly ambiguous. Ernst, Banks, and Bülthoff (2000) and Atkins, Fiser, and Jacobs (2001) found that subjects' estimates of visual depth relied more heavily on a visual cue (e.g., texture) when the cue was congruent with a haptic signal versus when it was incongruent with this signal. Atkins, Jacobs, and Knill (2003) reported that subjects recalibrate their interpretations of a visual stereo cue so that depth-from-stereo percepts are in greater agreement with depth-from-haptic percepts when visual and haptic signals are discrepant. Adams, Graf, and Ernst (2004) and Adams, Kerrigan, and Graf (2010) showed that visual-haptic experience can modify the visual system's "light-from-above" assumption used when observing images in which shading information to depth is ambiguous. Within the literature on artificial intelligence, there are relatively few articles studying how touch can educate vision. An exception is the work of Pinto, Gandhi, Han, Park, and Gupta (2016) showing that physical interactions (e.g., grasping, pushing, poking) can aid the acquisition of meaningful visual representations in a robot system.[1]

To date, the scientific literature on the affects of visual-haptic experience on visual learning is limited.[2] As indicated in the previous paragraph, experimental studies reported in this literature have focused on depth perception. Moreover, the literature lacks broad theoretical studies of what might be achieved by a learning system trained in a visual-haptic environment. The current paper aims to address this gap.

It does so through a set of computational investigations using artificial neural networks. Over the past several decades, researchers in the fields of cognitive science and neuroscience have used neural networks to provide insights into many aspects of human cognition, including perception, memory, language, reasoning, decision making, and action selection. Despite this history of success, the reasons as to why they often provide useful accounts of human perceptual and cognitive processing are poorly understood. For instance, there are many types of networks—networks can differ in terms of types of units, patterns of connectivity, training procedures, and many other factors—but researchers do not have a good understanding as to which types provide better versus worse accounts of human behavior. Similarly, many artificial neural networks lack biological detail, resembling biological neural networks only in seemingly coarse-scale ways. Despite this, several researchers have recently argued that these networks provide useful insights into neural processing, particularly within the visual system (Kriegeskorte, 2015; Wenliang & Seitz, 2018; Yamins & DiCarlo, 2016). Readers interested in the relationships between artificial and biological neural networks may see Churchland and Sejnowski (2017), Kriegeskorte and Golan (2019), Marblestone, Wayne, and Kording (2016), and Yamins and DiCarlo (2016).

The neural network models used in the research reported here have at least two properties that make them relevant to human perception. First, visual features in the models were extracted using a network with alternating layers of convolutional and pooling units (followed by fully connected layers) whose processing is reminiscent of processing in visual cortical areas. Previous researchers have found that processing in networks using convolution and pooling resembles biological visual processing in intriguing ways (Kriegeskorte, 2015; Wenliang & Seitz, 2018; Yamins & DiCarlo, 2016). Second, the models make explicit use of efficient data compression to learn compact representations of perceptual data items. Within the vision sciences, analyses of behavioral and neural responses from the perspective of *coding efficiency* have yielded important insights into perceptual processing (Barlow, 1961; Simoncelli & Olshausen, 2001; Sims, 2018).

In particular, our models make use of variational autoencoders (VAEs), neural networks that learn efficient data representations in an unsupervised manner (Kingma & Welling, 2014; Rezende, Mohamed, & Wierstra, 2014). β-VAEs are nonlinear models that are effective at learning latent (or hidden) variables underlying observed data. Relative to other nonlinear models that learn latent variables, an advantage of β-VAEs is that it is easy to adjust their information capacities, meaning that the latent representations they acquire can be weakly, moderately, or strongly constrained.

A VAE consists of two parts, an encoder that learns to map an input pattern to a compressed hidden or latent representation, and a decoder that learns to map the latent representation to an output pattern that approximates the input pattern. To achieve close approximations for novel test patterns, VAEs must acquire latent representations that contain information about the statistical regularities of the training patterns.

During training, VAEs adjust their parameter values to minimize the following objective function:

$$\mathcal{L}(\theta, \phi; \mathbf{x}, \mathbf{z}) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - \beta D_{KL}(q_\phi(\mathbf{z}|\mathbf{x}) \parallel p(\mathbf{z})) \quad (1)$$

where ϕ and θ denote the parameters of the encoder and decoder, respectively, $\mathbf{x}$ is a vector of input feature

values, and $\mathbf{z}$ is a vector of latent feature values. The right side of this equation has two terms. The first term is the expected log probability of input $\mathbf{x}$ given latent representation $\mathbf{z}$. It is often referred to as the *reconstruction error*. If, for example, each input feature has a normal distribution (given $\mathbf{z}$), then this term is the sum of squared error between the true input feature values and their estimated or reconstructed values. The second term is the Kullback-Leibler (KL) distance between the posterior distribution of latent representation $\mathbf{z}$ (after observing $\mathbf{x}$), denoted $q_\phi(\mathbf{z}|\mathbf{x})$, and its prior distribution, denoted $p(\mathbf{z})$. This term acts as a regularizer. It ameliorates potential *overfitting* (e.g., rote memorization of training data including the noise in these data), constraining the latent representation acquired during training by biasing the posterior distribution of this representation toward its prior distribution.

For VAEs, the coefficient β is set to one. β-VAEs are a variant of VAEs in which β can be set to any nonnegative value (Higgins et al., 2016). Consequently, a researcher can control the extent to which a β-VAE's acquired latent representation is regularized or constrained through the choice of β. Interestingly, both VAEs and β-VAEs can be characterized using rate-distortion theory (Alemi et al., 2018; Burgess et al., 2018), a branch of information theory that seeks to understand lossy compression in communication channels by quantifying relationships between rate (or information capacity) and distortion (or reconstruction error). Rate-distortion theory has been shown to provide good accounts of aspects of human visual perception (Bates, Lerch, Sims, & Jacobs, 2019; Sims, 2016; Sims, Jacobs, & Knill, 2012).

The connection between β-VAEs and rate-distortion theory plays a pivotal role in the results discussed below. Intuitively, when β is set to a large value, a β-VAE is highly constrained, thereby resembling a communication channel with low rate or capacity. Critically, the optimal outputs of a low-rate channel are typically summaries or abstractions of its inputs. This occurs because the channel has a low rate, and thus it cannot reconstruct all the fine-scale details of its inputs. The best it can do is to reconstruct its inputs' abstract or coarse-scale features.

Consider, for example, a scenario in which each input belongs to either category A or B. When receiving an input from category A, the optimal output of a low-rate channel will resemble the prototype for category A, and when receiving an input from B, the output will resemble B's prototype. That is, the optimal channel's outputs will reflect the categorical structure of its inputs. As demonstrated below, β-VAEs have a similar property. When sufficiently constrained (i.e., when β is set to a large value), β-VAEs learn the abstract or categorical structures of their training data.

Visual and haptic feature values

In our simulations, the visual and haptic feature values were derived from the See-and-Grasp data set.[3] This data set is based on novel objects known as Fribbles (Barry, Griffith, De Rossi, & Hermans, 2014; Hayward & Williams, 2000; Tarr, 2003; Williams, 1997). Fribbles are 3D, multipart, naturalistic objects with a categorical structure. The See-and-Grasp data set contains 891 Fribbles organized into three Fribble families, with four species in two of the families and three species in the remaining family, and 81 Fribbles in each species. Each Fribble contains a part known as its main body. Fribbles in the same family have a common main body. In addition to a main body, each Fribble has four slots, and one of three possible parts is attached at each slot. Fribbles in the same species use the same set of possible parts (four slots with three possible parts per slot = 81 Fribbles per species).

Columns 1–2, 3–4, and 5–6 of Figure 1 show sample images of Fribbles from the first, second, and third families, respectively, and from original and left-right flipped viewpoints or orientations. All objects were visually rendered from a 3/4-view so that object parts are visible and object shapes are easily perceived.

Each image consisted of 400×400 pixel values. The visual feature values for an image were generated in three steps. First, an image was resized and used as an input to a VGG deep neural network (Simonyan & Zisserman, 2015; we used the version of VGG-16 available in the Keras neural network library [Chollet, 2017]). The activation values of the hidden units at the last layer of this network's convolutional base were extracted. The output shape at this layer is $7 \times 7 \times 512$ meaning that there were 25,008 activation values per image. We used VGG because it shows good performance, having secured the first and second places in the 2014 ImageNet Challenge localization and classification tracks. In addition, its representations capture important aspects of people's image similarity ratings, often better than alternative deep neural networks (Peterson, Abbott, & Griffiths, 2018).

Next, we performed principal component analysis (PCA) on the VGG activation values for the entire set of images, and projected these values onto the 200 components on which the activation values showed the highest variance. These 200 components accounted for more than 97% of the variance in the activation data. Consequently, each image was represented by 200 visual feature values. Finally, each feature was normalized so that its values had a mean of zero and a variance of one.

To represent Fribbles in the haptic domain, the See-and-Grasp data set used the GraspIt! grasp simulator developed in the robotics community (Miller & Allen, 2004). GraspIt! contains a simulator of a human hand.

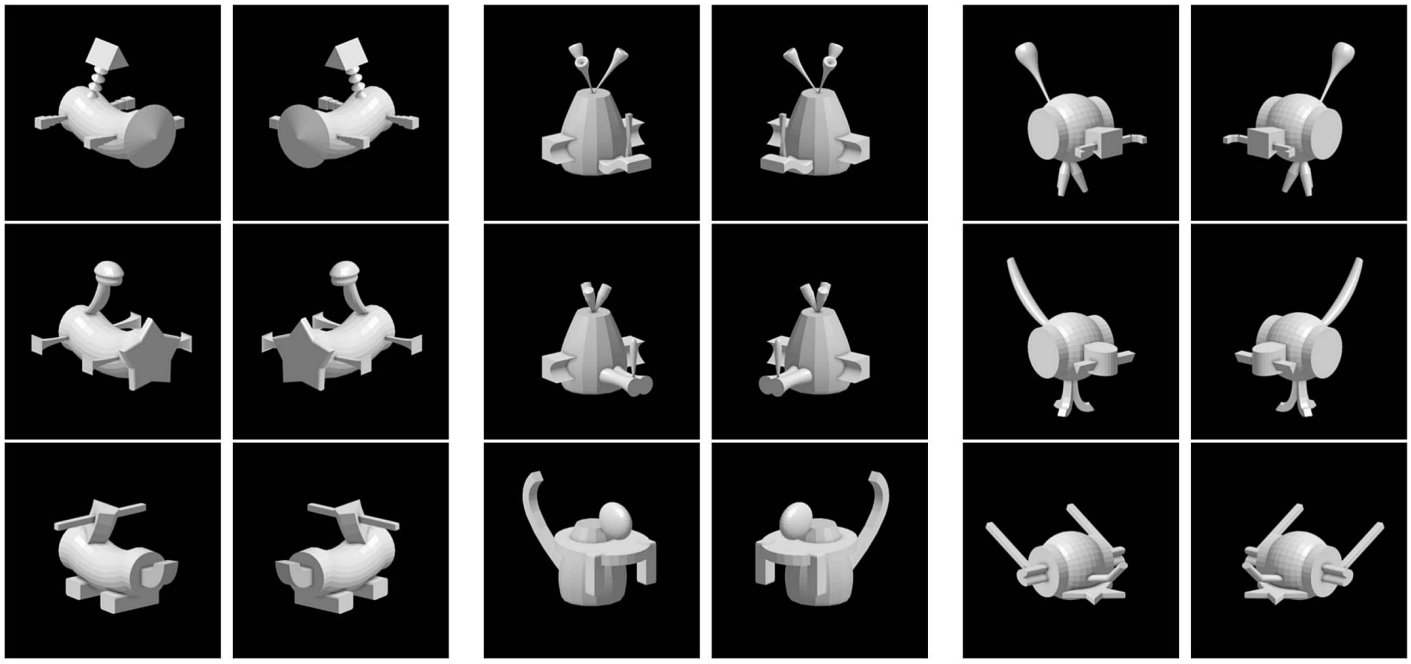

Figure 1. Columns 1–2, 3–4, and 5–6 show images (at original and flipped orientations) of Fribbles from the first, second, and third families, respectively.

When calculating the haptic features of a Fribble, the input to GraspIt! was the 3D shape model for the Fribble. Its output was a set of 16 joint angles of the fingers of the simulated human hand obtained when the hand "grasped" the object (the joint angles characterize the shape of the hand as the hand is grasping a Fribble). Grasps—or closings of the fingers around an object—were performed using GraspIt!'s AutoGrasp function. Figure 2 shows the simulated hand grasping an object at three orientations. In the See-and-Grasp data set, each object was grasped 24 times, each time from a different orientation (different orientations were generated by rotating an object eight times [each time by 45°] around the width, length, and depth axes). The use of multiple grasps can be regarded as an approximation to active haptic exploration. Consequently, a Fribble was haptically represented by a vector with 384 elements (16 joint angles per grasp $\times$ 24 grasps).

In our simulations, we performed PCA on the haptic vectors for the entire set of Fribbles, and projected the vector elements onto the 200 components on which the haptic data showed the highest variance. These 200 components accounted for more than 99% of the variance in the haptic data. Consequently, each Fribble was represented by 200 haptic feature values. Each feature was then normalized so that its values had a mean of zero and a variance of one.

The data set used here has several advantages and disadvantages. Our primary reason for using it is that it includes both visual and haptic features for naturalistic objects with naturalistic organizations. One such

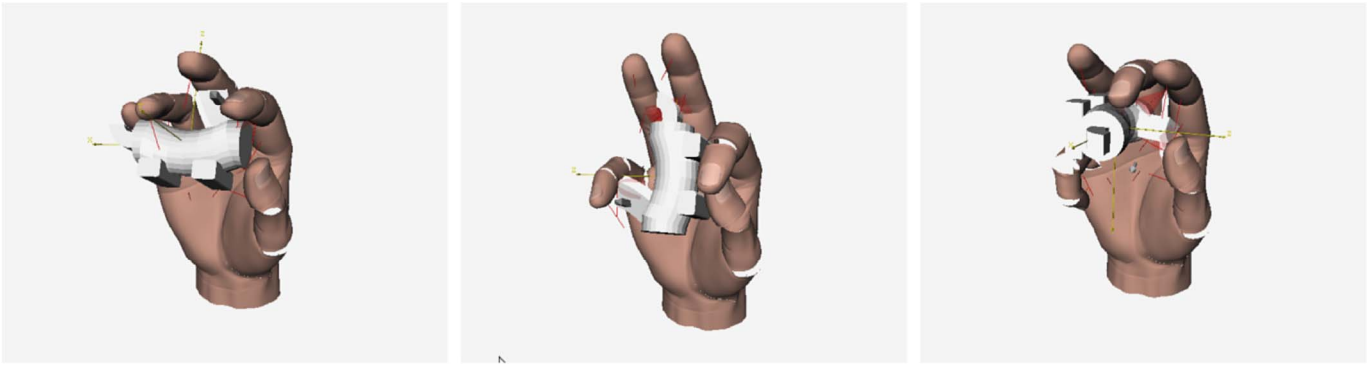

Figure 2. GraspIt! simulates a human hand. Here the hand is grasping an object at three different orientations. Reprinted from Erdogan, Yildirim, and Jacobs (2015).

organization is the objects' categorical structure: objects are exemplars from categories. Another organization is that objects are part-based where objects share a discrete set of easily identifiable parts. As discussed below, our simulation results evaluate the "goodness" of models' acquired object representations based on how well these representations reflect objects' categorical and part-based organizations. To our knowledge, this is the only publicly available data set with these desirable properties.

Of course, the data set has disadvantages too. First, because there are 891 objects and each object was visually rendered from two orientations, the data set has 1,782 data items. That is, the data set is small, especially by the standards of artificial intelligence researchers who often use data sets with millions of items. Second, although objects are naturalistic, they are not natural. For example, the results reported here were obtained with objects with discrete parts. Different results might be found with other types of objects such as "blobby" amoeboids. Third, data items reflect an admittedly extreme situation. We aim to evaluate the usefulness of multisensory training for visual learning in the most favorable conditions because if multisensory training is not beneficial in these conditions, then it will not be beneficial in other conditions. Consequently, data items come from a situation in which visual signals are relatively impoverished, whereas haptic signals are rich. In particular, the visual feature values for a data item come from a single static image, and thus are orientation-dependent. The haptic feature values, in contrast, come from multiple grasps at multiple orientations, and thus are orientation-independent. This situation arises when a person looks at an object and grasps the object multiple times from multiple orientations. Although biased, this is a natural situation, experienced by many people on many occasions. Nonetheless, different results would be found with other types of situations. Future work will need to investigate other situations.

Simulation Details

We simulated two β-VAE models, referred to as Model V-H (top row of Figure 3) and Model V (bottom row).[4] Model V-H was trained with both visual and haptic feature values, whereas Model V was trained with visual feature values only. At first glance, it may seem intuitive that Model V-H should outperform Model V—after all, it is trained with more sensory information. However, this is not a foregone conclusion. First, Model V-H is trained with both visual and haptic features but, as discussed below, it is tested with visual features only. It may be that the difference between training and test conditions leads to poor performance by Model V-H during test. In addition, there is an interaction between a model's information capacity and its performance. As demonstrated below, Model V-H can take advantage of its extra sensory information, but the advantages of multisensory training are often greatest when the model's capacity is constrained.

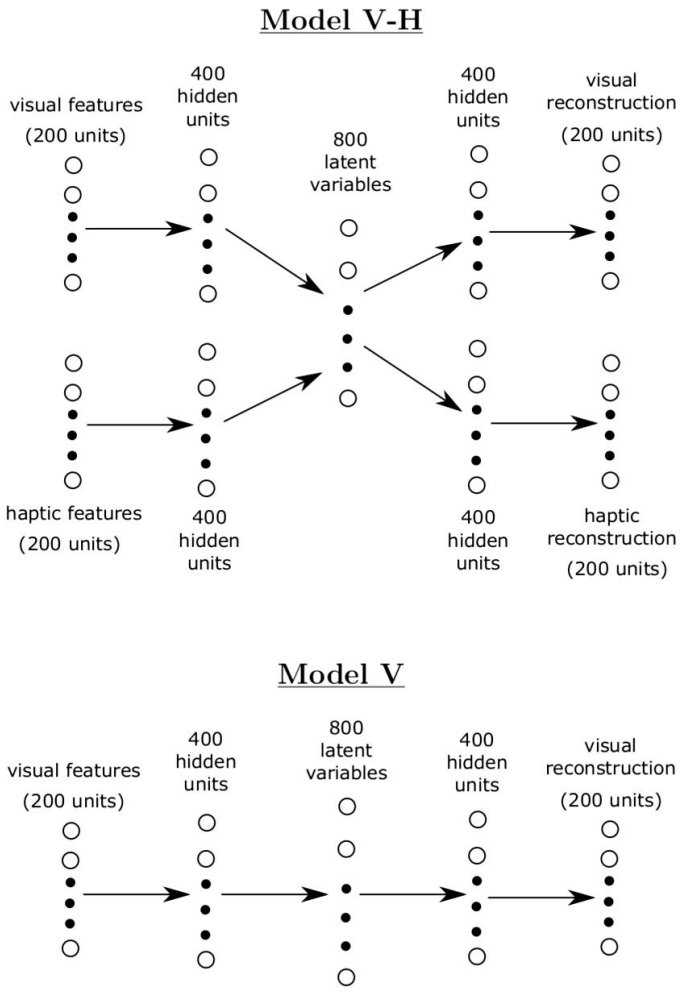

Figure 3. Models V-H (top) and V (bottom). Unfilled circles are individual neural units, sequences of three filled circles are ellipses indicating that some units are omitted from an illustration, and arrows denote a full set of connections between units.

Models were implemented using the Keras software package (Chollet, 2017).[5] Simulations used β values of 0.01, 1.0, 2.5, 5.0, 10.0, and 20.0. Hidden units of a model used the hyperbolic tangent (tanh) activation function. Units comprising the latent representation and the output units used a linear activation function. The prior distribution of the latent representation was a unit normal distribution (mean equals the zero vector; covariance equals the identity matrix).

Training lasted 2,000 epochs. During training, optimization of the weight values was performed using stochastic gradient descent (batch size = 64 data items; learning rate = 0.01). Training and testing were conducted using 10-fold cross-validation (Hastie, Tibshirani, & Friedman, 2009). That is, the set of 1,782 data items was randomly divided into 10 subsets (with the constraint that each subset contained roughly the same number of Fribbles from each species). On each fold, nine subsets were used for training and the remaining subset was used for testing. This was repeated for 10 folds such that each subset was used exactly once for testing.

Simulation results

Visualizations of latent representations

To gain insight into the latent representations of data items acquired during training, we visualized (the mean values of) these representations in two dimensions using t-distributed stochastic neighbor embedding (t-SNE), a nonlinear dimensionality-reduction technique that is often useful for visualizing high-dimensional data (van der Maaten & Hinton, 2008).[6] We used the latent representations obtained on the training data items. However, haptic feature values of items were set to zero for Model V-H. In other words, we visualized the latent representations based solely on visual feature values for both Model V-H and Model V.

The results for the first fold are shown in Figure 4. The top and bottom portions show the results for Models V-H and V, respectively. In the top row of each portion, data items are colored based on the Fribble family of each item (thus, there are three colors). The six plots in a row show the results corresponding to the six values of β (in increasing order from $\beta = 0.01$ on the left to $\beta = 20.0$ on the right).

The results are revealing. Consider, for example, the top row for Model V-H. The plots in this row clearly show that, as the model was more constrained (i.e., moving in the row from left to right), the model learned that there are 11 species of Fribbles (as indicated by the 11 clusters in the top right plot). In other words, the model learned about the categorical structure of the Fribbles. Impressively, it learned about this categorical structure despite the fact that it was never explicitly given training information regarding this structure. In contrast, Model V learned about the cross-product of Fribbles species membership and visual orientation, but only when the model was relatively unconstrained (see top left plot in bottom portion of Figure 4 that has 22 clusters [11 species × 2 orientations]). Relative to Model V, it seems as if Model V-H learned about species membership in an orientation-insensitive manner, though only when the model was constrained.

The hypothesis that Model V-H learned an orientation-insensitive latent representation is strengthened when one considers the bottom rows of each portion of the figure. In these rows, items are colored based on the visual orientation of each item (thus, there are two colors). Consider, for example, the right-most plot of the bottom row for Model V-H. Clusters in this plot for the two visual orientations show extensive overlap suggesting that Model V-H learned similar latent representations for Fribbles regardless of whether Fribbles were visually rendered from original or flipped orientations. This occurred when the model was relatively constrained, but not when it was unconstrained. In contrast, Model V never learned an orientation-insensitive latent representation. The finding that Model V-H learned orientation-insensitive representations presumably arose because this model received as input orientation-dependent visual features that were paired with orientation-independent haptic features. A more direct test of the orientation sensitivity of the models is presented below.

Visual and haptic reconstructions

Using the test items in each fold of the cross-validation procedure, we calculated the sum of squared error (SSE) of each model's visual reconstructions based solely on items' visual feature values (for Model V-H, haptic feature values were set to zero). The results are shown in the left graph of Figure 5. Model V had smaller errors than Model V-H, especially when it was relatively unconstrained (i.e., trained with small values of β), indicating that Model V learned more about the fine-scale visual structure of items.[7] The superior SSE performance of Model V was expected because Model V-H was trained to learn item's visual and haptic structures, whereas Model V was trained to learn only the visual structure, and because Model V-H was trained with haptic feature values but was tested in the absence of these values.

We also evaluated whether Model V-H could reconstruct items' haptic feature values based solely on their visual feature values. As illustrated in the right graph of Figure 5, its haptic reconstructions were only moderately worse than its visual reconstructions. This is interesting because it has been hypothesized that people are able to make cross-modal sensory predictions, at least at a coarse level of detail. For example, Smith and Goodale (2015) showed human subjects images of objects while subjects were in an fMRI scanner. They found that they could decode the categories of viewed objects at above-chance levels from the voxel activations of subjects' early regions of

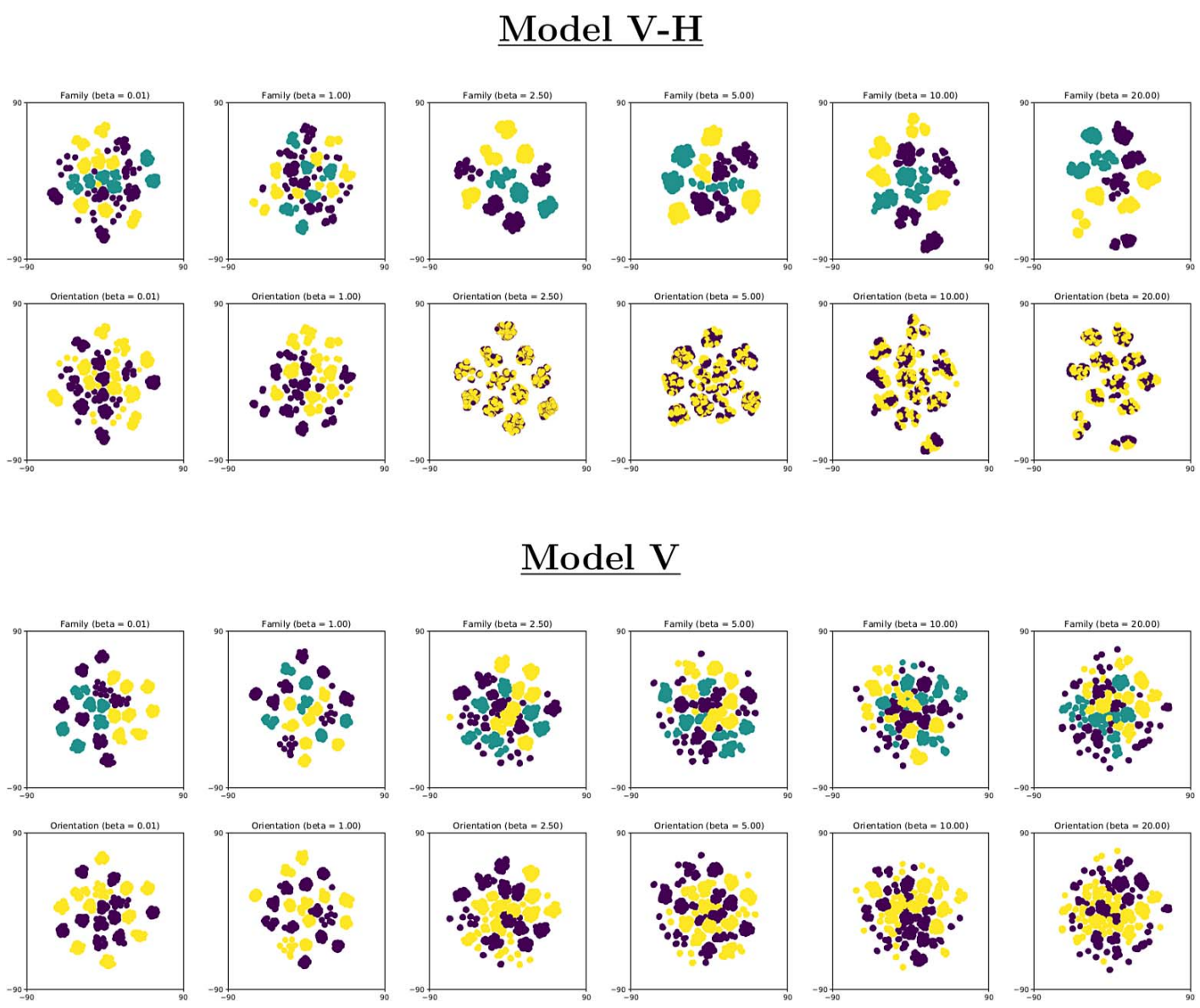

Figure 4. t-SNE visualizations of (the mean values of) the latent representations of Models V-H (top rows) and V (bottom rows). Data were collected during the first fold of the 10-fold cross-validation procedure. Data items were training items that contained visual feature values only (for Model V-H, haptic feature values were set to zero). The six columns show visualizations when β was set to 0.01, 1.0, 2.5, 5.0, 10.0, and 20.0, respectively. The colors in rows 1 and 3 indicate the Fribble family membership of data items. The colors in rows 2 and 4 indicate whether the visual image of a data item was in original or flipped orientation.

somatosensory cortex, though it was not possible to decode the specific objects that were viewed. This result suggests that subjects used objects' visual feature values to predict their tactile or haptic feature values, and that these predicted haptic feature values were sufficient to estimate objects' categories.

Classifications of objects' structures

To further understand models' learning performances, we examined how well their mean latent representations could predict different aspects of Fribbles' underlying structures. Four classification tasks were considered:

- Classify the family of the Fribble associated with each data item. There are three Fribble families.
- Classify the species of the Fribble associated with each data item. There are 11 Fribble species.
- Classify the identity of the Fribble associated with each data item. There are 891 different Fribbles.
- Classify the parts comprising the Fribble associated with each data item. Ignoring a Fribbles main

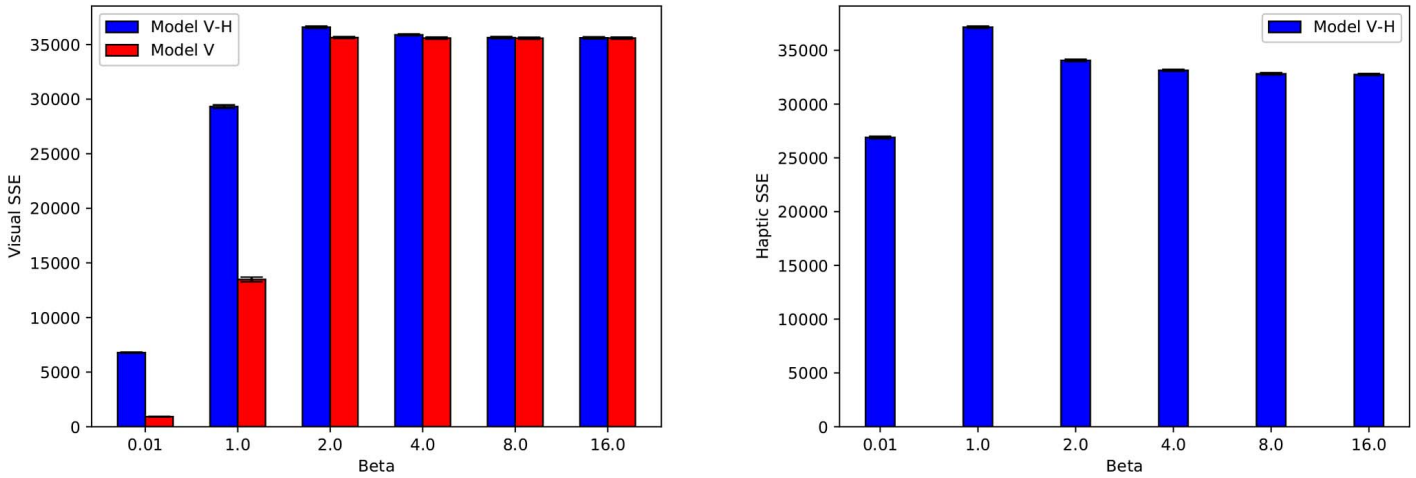

Figure 5. (Left) Sum of squared errors (SSEs) for visual reconstructions based on visual feature values of test items as a function of the value of β. Blue and red bars are for Models V-H and V, respectively. Error bars (often too small to see) indicate the standard errors of the means across the 10 folds of the cross-validation procedure. (Right) SSEs for haptic reconstructions based on visual feature values of test items for Model V-H. Data for Model V are not shown because this model does not know about haptic feature values.

body (which defines a Fribble's family), each Fribble is composed of four parts. Across all species, there are 132 possible parts, and thus this task consisted of 132 binary classification subtasks.[8]

Tasks were performed using a linear support vector machine (SVM). To generate inputs for this SVM, we performed PCA on the set of mean latent representations of data items, and projected these representations onto the 40 principal components on which the representation values had the largest variance. Data items were visual feature values of test items from the cross-validation training procedure (as above, haptic feature values were set to zero for Model V-H so that Model V-H and Model V performances are based solely on visual feature values).

Classification performances on the four tasks are shown in the four graphs of Figure 6. Taken as a whole, Model V-H clearly outperformed Model V. Model V often performed best when it was relatively unconstrained. It seems that when it was moderately or highly constrained, it may have learned abstract latent representations, but the abstractions codified in these representations did not necessarily correspond to the underlying organizations of the Fribbles. In contrast, Model V-H learned abstract latent representations that better corresponded to Fribbles' underlying organizations.

Orientation insensitivity

The visualizations of latent representations discussed above (Figure 4) suggested that Model V-H, but not Model V, learned representations that were relatively insensitive to an object's orientation in an image. To more directly address this issue, we did the following. For each test item, we calculated a model's latent representation based solely on the item's visual feature values (as above, haptic feature values were set to zero). Call this the *target* representation. We also calculated the latent representations for every other item (regardless of whether the other item was a training item or a test item). Call these the *probe* representations. Next, we correlated the target representation with each of the probe representations, and found the data items corresponding to the probes with the five highest correlations. Call this set the *most similar* data items. Lastly, if the most-similar data items included an item depicting the same object as the test item but with a different orientation, then the test item was regarded as being correctly represented.

The results are shown in Figure 7. When $\beta = 2.5$, Model V-H learned latent representations that were fully orientation insensitive. For other moderate and large values of β, this model learned latent representations that were largely orientation insensitive. In contrast, Model V never learned orientation-insensitive latent representations.

Discussion

In the field of machine learning, it is typically thought that a learning system will perform best when training and test data items have the same statistical properties. For example, a system that will be tested with visual images should be trained with visual images.

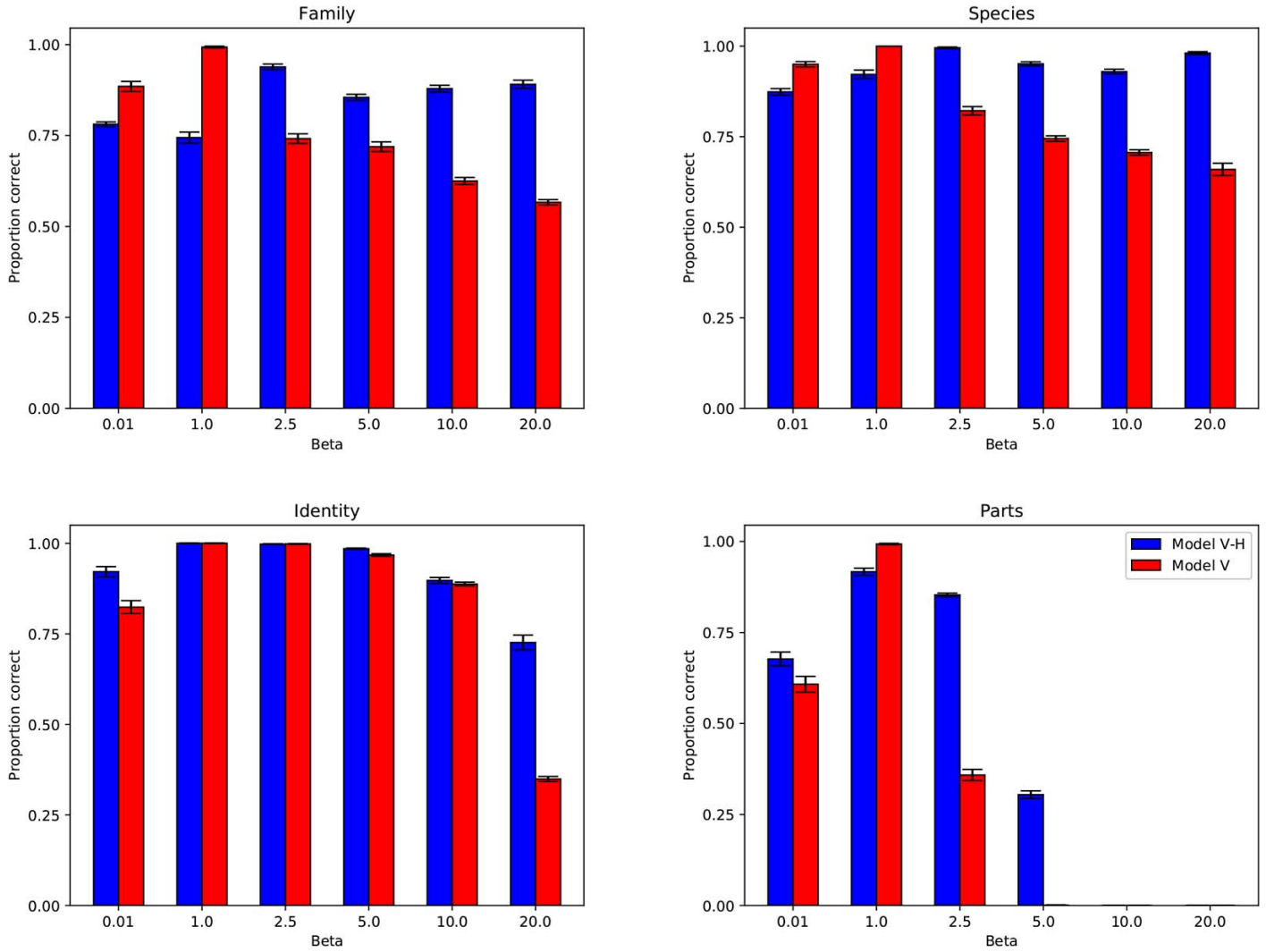

Figure 6. Classification performances (proportion correct) based on visual feature values of test items as a function of the value of β. Classification tasks are Fribble family membership (top left), Fribble species membership (top right), individual Fribble identity (bottom left), and parts of a Fribble (bottom right). Blue and red bars are for Models V-H and V, respectively. Error bars indicate the standard errors of the means.

Here, we have explored an exception to this rule. We have found that a system tested with images obtains important benefits when it is trained in a multisensory environment containing both visual and haptic sensory features (Model V-H) relative to when it is trained in a unisensory environment containing visual features only (Model V). In particular, our results demonstrate that a system trained in a visual-haptic environment (in which visual, but not haptic, signals are orientation-dependent) tends to learn visual representations containing useful abstractions, such as the categorical structure of objects, and also learns representations that are less sensitive to imaging parameters, such as viewpoint or orientation, that are irrelevant for object recognition or classification tasks.

Our results are pertinent to both cognitive scientists with interests in human visual perception and computer scientists with interests in computer vision. Both sets of scientists often study perceptual learning in unisensory environments containing visual features only. Consequently, these scientists may be overestimating the difficulties associated with important perceptual learning problems. Although multisensory perception has its own challenges, our results demonstrate that perceptual learning can become easier when it is considered in a multisensory context (Shams & Seitz, 2008).

In this paper, we have been careful not to claim that Model V-H, the system trained in a multisensory environment, is strictly better for visual learning than Model V, the system trained in a unisensory environment. Indeed, there is an interaction between the

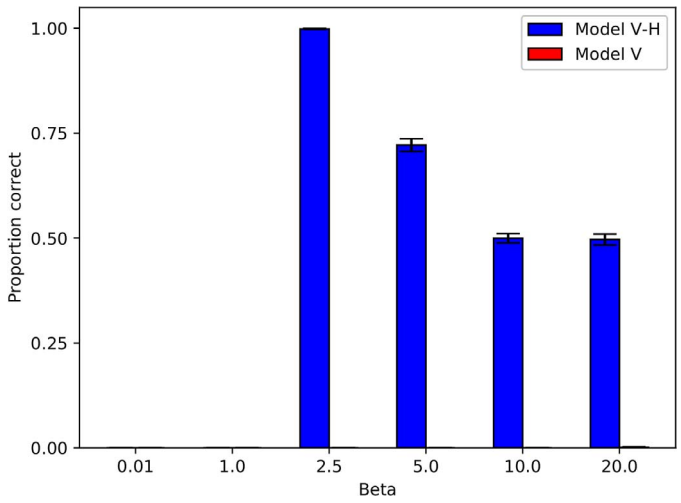

Figure 7. Orientation-insensitivity performances (proportion correct) based on visual feature values of test items as a function of the value of β (see text for explanation). Blue and red bars are for Models V-H and V, respectively (red bars are barely visible due to near-zero values). Error bars indicate the standard errors of the means.

information capacity of a system, the nature of the sensory information that the system receives, and the detail of visual information that a system learns. Many of our results indicate that Model V is better at learning fine-scale, two-dimensional properties of visual images when it has a large information capacity, whereas Model V-H is better at learning coarse-scale, three-dimensional properties of objects depicted in images when it has a small information capacity. If so, this suggests that agents (biological or artificial) should contain multiple learning systems, varying in capacity (from low to high capacity) and sensory input (from unisensory to multisensory). This would allow an agent to learn a hierarchy of visual representations ranging from representations that codify fine-scale image features to those that codify coarse-scale abstractions. Our results demonstrate that information from multiple sensory modalities can guide learning systems toward useful visual abstractions.

Keywords: visual learning, multisensory perception, computational modeling

Acknowledgments

We thank Chris Bates for many helpful discussions. This work was supported by National Science Foundation research grants (BCS-1400784, BCS-1824737, IIS-1741472, and IIS-1813709).

Commercial relationships: none.
Corresponding author: Robert A. Jacobs.
Email: robbie@bcs.rochester.edu.
Address: Department of Brain and Cognitive Sciences, University of Rochester, Rochester, NY, USA.

Footnotes

[1] Illustrating the rarity with which AI researchers have studied how touch can educate vision, Pinto et al. (2016) wrote: "While there has been significant work in the vision and robotics community to develop vision algorithms for performing robotic tasks such as grasping, to the best of our knowledge this is the first effort that reverses the pipeline and uses robotic tasks for learning visual representations" (p. 2).

[2] Readers interested in the effects of visual-auditory experience on visual learning should see Shams and Seitz (2008).

[3] An early version of this data set was described in Yildirim and Jacobs (2013); see also Erdogan, Chen, Garcea, Mahon, and Jacobs (2016), and Erdogan, Yildirim, and Jacobs (2015). The current version is available at https://zenodo.org/record/3266251#.XRuZ4pNKhYc (https://doi.org/10.5281/zenodo.3266251).

[4] Strictly speaking, the networks shown in Figure 3 are not fully accurate. In both models, the latent representation consisted of 800 random variables independently sampled from normal distributions. Let z_i denote the i^{th} latent variable, and let μ_i and σ_i^2 denote its mean and variance, respectively. The output of the encoder portion of a model consisted of 800 activation values corresponding to $\{\mu_i\}_{i=1}^{800}$ and 800 activation values corresponding to$\{\log \sigma_i^2\}_{i=1}^{800}$. The $\{z_i\}_{i=1}^{800}$ served as inputs to the decoder portion of a model.

[5] Sample code can be found at https://zenodo.org/record/3266341#.XRuyLpNKhYc (https://doi.org/10.5281/zenodo.3266341)

[6] Although t-SNE can be "brittle", all the results reported here were obtained using the "scikit-learn" implementation (Pedregosa et al., 2011) with its default parameter settings.

[7] Rendering the visual reconstructions as images is not easily accomplished due to difficulties with inverting the nonlinear VGG network (i.e., mapping from VGG features to images). As a sanity check, we repeated our experiments using visual features obtained by applying PCA directly to images. SSE performances were qualitatively similar to those plotted in Figure 5. Images obtained from the visual reconstructions were also as expected. For example, images generated from Model V's reconstructions at

high capacity were very similar to original images, whereas images generated from Model V-H's reconstructions at low capacity resembled blurry averages of original images.

[8] If a part did not exist in a set of test items, then the subtask corresponding to this part was omitted.

References

Adams, W. J., Graf, E. W., & Ernst, M. O. (2004). Experience can change the 'light-from-above' prior. *Nature Neuroscience*, *7*, 1057–1058.

Adams, W. J., Kerrigan, I. S., & Graf, E. W. (2010). Efficient visual recalibration from either visual or haptic feedback: The importance of being wrong. *Journal of Neuroscience*, *30*, 14745–14749.

Alemi, A. A., Poole, B., Fischer, I., Dillon, J. V., Saurous, R. A., & Murphy, K. (2018). Fixing a broken ELBO. *arXiv*:1711.00464.

Atkins, J. E., Fiser, J., & Jacobs, R. A. (2001). Experience-dependent visual cue integration based on consistencies between visual and haptic percepts. *Vision Research*, *41*, 449–461.

Atkins, J. E., Jacobs, R. A., & Knill, D. C. (2003). Experience-dependent visual cue recalibration based on discrepancies between visual and haptic percepts. *Vision Research*, *43*, 2603–2613.

Barlow, H. B. (1961). Possible principles underlying the transformation of sensory messages. In W. A. Rosenblith (Ed.), *Sensory communication* (pp. 217–234). Cambridge, MA: MIT Press.

Barry, T. J., Griffith, J. W., De Rossi, S., & Hermans, D. (2014). Meet the Fribbles: Novel stimuli for use within behavioural research. *Frontiers in Psychology*, *5*:103, 1–8.

Bates, C. J., Lerch, R. A., Sims, C. R., & Jacobs, R. A. (2019). Adaptive allocation of human visual working memory capacity during statistical and categorical learning. *Journal of Vision*, *19*(2):11, 1–23, https://doi.org/10.1167/19.2.11. [PubMed] [Article]

Berkeley, G. (1709 /1910). *An essay towards a new theory of vision*. London: Dutton.

Burgess, C. P., Higgins, I., Pal, A., Matthey, L., Watters, N., Desjardins, G., & Lerchner, A. (2018). Understanding disentangling in β-VAE. *arXiv*: 1804.03599.

Churchland, P. S. & Sejnowski, T. J. (2017). *The Computational Brain (25th Anniversary Edition)*. Cambridge, MA: MIT Press.

Chollet, F. (2017). *Deep learning with Python*. Shelter Island, NY: Manning Publications.

Erdogan, G., Chen, Q., Garcea, F. E., Mahon, B. Z., & Jacobs, R. A. (2016). Multisensory part-based representations of objects in human lateral occipital cortex. *Journal of Cognitive Neuroscience*, *28*, 869–881.

Erdogan, G., Yildirim, I., & Jacobs, R. A. (2015). From sensory signals to modality-independent conceptual representations: A probabilistic language of thought approach. *PLoS Computational Biology*, *11*(11): e1004610.

Ernst, M. O., Banks, M. S., & Bülthoff, H. H. (2000). Touch can change visual slant perception. *Nature Neuroscience*, *3*, 69–73.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). New York: Springer.

Hayward, W. G., & Williams, P. (2000). Viewpoint dependence and object discriminability. *Psychological Science*, *11*, 7–12.

Held, R., Ostrovsky, Y., de Gelder, B., Gandhi, T., Ganesh, S., Mathur, U., & Sinha, P. (2011). The newly sighted fail to match seen with felt. *Nature Neuroscience*, *14*, 551–553.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., ... Lerchner, A. (2017). β-VAE: Learning basic visual concepts with a constrained variational framework. *Proceedings of the 2017 International Conference on Learning Representations*.

Kingma, D. P., & Welling, M. (2014). Auto-encoding variational Bayes. *arXiv*:1312.6114.

Kriegeskorte, N. (2015). Deep neural networks: A new framework for modeling biological vision and brain information processing. *Annual Review of Vision Science*, *1*, 417–446.

Kriegeskorte, N., & Golan, T. (2019). Neural network models and deep learning. *Current Biology*, *29*, R231–236.

Marblestone, A. H., Wayne, G., & Kording, K. P. (2016). Towards an integration of deep learning and neuroscience. *arXiv*:1606.03813.

Miller, A. & Allen, P. K. (2004). GraspIt!: A versatile simulator for robotic grasping. *IEEE Robotics and Automation Magazine 11*, 110–122.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Peterson, J. C., Abbott, J. T., & Griffiths, T. L. (2018).

Evaluating (and improving) the correspondence between deep neural networks and human representations. *Cognitive Science*, *42*, 2648–2669.

Piaget, J. (1952). *The origins of intelligence in children*. New York: International Universities Press.

Pinto, L., Gandhi, D., Han, Y., Park, Y.-L., & Gupta, A. (2016). The curious robot: Learning visual representations via physical interactions. *arXiv*: 1604.01360.

Rezende, D. J., Mohamed, S., & Wierstra, D. (2014). Stochastic backpropagation and approximate inference in deep generative models. *arXiv*:1401.4082.

Shams, L. & Seitz, A. R. (2008). Benefits of multisensory learning. *Trends in Cognitive Sciences*, *12*, 411–417.

Simoncelli, E. P., & Olshausen, B. A. (2001). Natural image statistics and neural representation. *Annual Review of Neuroscience*, *24*, 1193–1216.

Simonyan, K. & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. *Proceedings of the 2015 International Conference on Learning Representations*.

Sims, C. R. (2016). Rate-distortion theory and human perception. *Cognition*, *152*, 181–198.

Sims, C. R. (2018). Efficient coding explains the universal law of generalization in human perception. *Science*, *360*, 652–656.

Sims, C. R., Jacobs, R. A., & Knill, D. C. (2012). An ideal observer analysis of visual working memory. *Psychological Review*, *119*, 807–830.

Smith, F. W., & Goodale, M. A. (2015). Decoding visual object categories in early somatosensory cortex. *Cerebral Cortex*, *25*, 1020–1031.

Tarr, M. J. (2003). Visual object recognition: Can a single mechanism suffice? In M. A. Peterson & G. Rhodes (Eds.), *Perception of faces, objects, and scenes: Analytic and holistic processes* (pp. 177–211). New York: Oxford University Press.

van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*, 2579–2605.

Wenliang, L. K. & Seitz, A. R. (2018). Deep neural networks for modeling visual perceptual learning. *Journal of Neuroscience*, *38*, 6028–6044.

Williams, P. (1997). *Prototypes, exemplars, and object recognition* (Unpublished doctoral dissertation). Department of Psychology, Yale University.

Yamins, D. L. K., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, *19*, 356–365.

Yildirim, I., & Jacobs, R. A. (2013). Transfer of object category knowledge across visual and haptic modalities: Experimental and computational studies. *Cognition*, *126*, 135–148.